Clinically relevant sample size calculations for trials

Papageorgiou, Spyridon N

Abstract: An orthodontist attends the scientific orthodontic congress and passing by the company exhibitions receives a free sample of a novel orthodontic archwire introduced in the market for initial alignment/levelling with favourable performance over conventional superelastic nickel-titanium archwires. This novel archwire should increase according to the manufacturer's claims alignment efficacy and halve the pain for the patient. After using the free sample wire on a patient and observing that the patient felt little to no pain during the alignment/levelling phase, the orthodontist decides to venture into unknown territory and find out if this holds true for most of her patients by doing a randomised clinical trial in her practice.

DOI: https://doi.org/10.1177/1465312519848757

STATISTICAL CORNER

Clinical relevant sample size calculations for trials

Spyridon N. Papageorgiou

Clinic of Orthodontics and Pediatric Dentistry, Center of Dental Medicine, University of Zurich, Zurich, Switzerland

ORCID Spyridon N. Papageorgiou http://orcid.org/0000-0003-1968-3326

CONTACT Spyridon N. Papageorgiou Clinic of Orthodontics and Pediatric Dentistry, Center of Dental Medicine, University of Zurich, Plattenstrasse 11, Zurich CH 8032, Switzerland; snpapage@gmail.com.

Words in text: 1894

Disclosure statement

No potential conflict of interest was reported by the author.

Theoretical scenario

An orthodontist attends the scientific orthodontic congress and passing by the company exhibitions receives a free sample of a novel orthodontic archwire introduced in the market for initial alignment/levelling with favourable performance over conventional superelastic nickel-titanium archwires. This novel archwire should increase according to the manufacturer's claims alignment efficacy and halve the pain for the patient. After using the free sample wire on a patient and observing that the patient felt little to no pain during the alignment/levelling phase, the orthodontist decides to venture into unknown territory and find out if this holds true for most of her patients by doing a randomised clinical trial in her practice.

She opts for the simplest trial design possible and plans to do a parallel randomised trial with two groups: one receiving the novel (hereon 'experimental') archwire and the other receiving a conventional superelastic nickel-titanium (hereon 'control') archwire. She plans to assess patient-reported pain at its peak 24 hours after insertion of the first archwire using a Visual Analogue Scale (VAS) with a 100-mm horizontal line. Patients mark felt pain intensity between 'no pain' on the its left side and 'worst pain possible' on the right side; this is a valid and reliable tool for pain assessment. She also performs an a priori sample size calculation to decide how many patients she would need to recruit for her trial to be adequately powered to identify an existing difference in pain between the experimental and control archwire, if such a difference really does exist. The sample size calculation is based on an expected pain for the control archwire of 28.8 mm in VAS (standard deviation of 11.0 mm in VAS) according to a previous study (Sandhu and Sandhu, 2013) the author judged to be of similar setting and conduct with her own. Expecting to find, according to the manufacturer's claims, a 50% reduction in pain using a t-test for independent samples, she sets α (type I "false positive" error) at 5% and power (1 minus β [type II "false negative" error]) at 80%, as is many times done (MacFarlane, 2003). This calculation gives her a total sample of 22 patients needed (11 in each group), to which she adds another 15% to account for possible losses to follow-up and ends up recruiting 26 patients in total (13 in each group).

After collecting and analysing the trial's data, she finds to her astonishment no statistically significant or clinically relevant difference in pain during leveling and aligning between the two archwires (mean difference of 17%; $P>0.05$).

Which of the following statements are true, if any?

(A) As the trial included a sample size calculation and found no significant difference, we can be confident that no difference exists between the two archwire in reality.

(B) The sample size calculation that was performed ensures that this trial is adequately powered to identify any existing difference between the two archwires.

(C) Being too optimistic during sample size calculation can lead to a trial being underpowered to detect a clinically relevant effect.

(D) It is always prudent to include as many patients as possible in a clinical trial.

Discussion

Sample size calculations are based on several a priori assumptions made by the researcher, as has been previously shown (MacFarlane 2003; Pandis and Cobourne, 2015). These include among others the anticipated mean response in the control and the experimental groups and their standard deviations, which are needed to calculate the minimal clinically important difference (MCID) to be detected from the trial. As the name implies, this is the smallest treatment effect (i.e. difference between the two groups) that the trial should be able to identify, if such an effect really exists. This does not mean necessarily that a treatment effect (i.e. difference between the two groups) exists. Only that if such a treatment exists and is equal or larger than the MCID that an adequately powered trial has good chances to find this. So statement (A) is wrong.

Clinical trials should then have enough power to identify existing differences between groups being same or larger than the MCID, but they would probably be underpowered to identify an existing difference that is smaller than the MCID. Therefore, it is imperative that careful consideration is given in determining the expected MCID of a treatment and this has to be based on both clinical and statistical grounds. For one, an MCID should be large enough to affect clinical practice or the whole point of making a clinical trial becomes void. Additionally, MCIDs should be realistic enough, which means that they should correspond to the intervention's clinical performance that can be expected according the intervention's technical and biological characteristics. Insight on the expected average performance of a new treatment can also be gleamed by looking at the observed effects of similar interventions in other trials. Setting an overly optimistic or overly pessimistic MCID – be it intentional or not – can have dire consequences for the feasibility,

duration, costs, and scientific value of a clinical trial, as it influences the results of the sample size calculation.

This can be seen in Table 1, where sample size calculations for a series of different MCIDs are given, while the anticipated mean pain response in the control group is kept the same. Differences in pain between the two groups are given both in terms of relative reduction in mm on the VAS scale (unstandardized effect size) and as Cohen's d (Cohen, 1969). Cohen's d is a standardised effect size, which is simply the difference between the two means divided by their pooled standard deviation and has no measurement units. This has the advantage that it can be used across different outcomes or disciplines to compare the magnitude of an effect, while at the same time having the disadvantage of not being clinically intuitive. Cohen introduced arbitrary cut-offs for the interpretation of d of such as small (d=0.2), medium (d=0.5) and large (d=0.8), but at the same time urged for strong caution that "this is an operation fraught with many dangers". The Cohen's d is used in this scenario to illustrate in simple terms the difference in magnitude and the impact it can have on sample size calculations. The original trial presented in the theoretical scenario set adopted a relative reduction of 50% or a very large Cohen's d of 1.3 (i.e. a difference of 1.3 times the standard deviation) and resulted in a needed sample of 22 patients. One can see that as the MCID becomes smaller the needed sample to keep the same power increases drastically: the needed sample for a large, medium, or small effect according to Cohen would be 52 patients (d=0.8), 128 patients (d=0.5), and 352 patients (d=0.3) (pink dotted line in Figure 1).

Table 1. Results of sample size calculations to identify an existing difference in pain 24 hours after archwire insertion with a t-test for independent samples, α at 5%, and power (1-β) at 20%.

Trial	Control group Mean*	Experimental group Mean	Relative reduction	Cohen's d	Total sample needed
A	28.8 mm	14.4 mm	-50%	1.3	22
B	Same	17.7 mm	-39%	1.0	34
C	Same	20.0 mm	-31%	0.8	52
D	Same	23.3 mm	-19%	0.5	128
E	Same	25.5 mm	-11%	0.3	352

* the control group mean and standard deviation were taken from the previous trial of Sandhu and Sandhu (2013). We assume that the control and experimental group would have the same standard deviation.

Both Table 1 and Figure 1 should be put, if possible, on the second or third page of the type-setted paper

Being too optimistic at the trial's start and expecting that a very large MCID of d=1.3 – while much smaller differences of d between 0.3-0.5 are found in reality – results in clinical trial with a very small sample. This means that the present trial might be severely underpowered to detect an existing difference in pain that is

less than 50% and statement (B) is wrong. Such differences 31% (d=0.8) or even 19% (d=0.5) might still be clinically relevant to the patient or the orthodontist, especially if the novel archwire comes with little to no extra costs or adverse effects compared to the conventional archwire. The current trial would have only 43% power to detect a d=0.8 and 20% power to detect a d=0.5. Or expressed as type II errors (β), the current trial has a 57% or 80% chance to miss identifying a clinically relevant effect that might exist. Therefore, it might well be that a difference in the pain response between the two archwire exists, but this trial could not find it, and statement (C) is correct.

Notwithstanding this, being too pessimistic of an expected treatment effect and setting a very small MCID of d=0.3 would lead the orthodontist to recruit at least 352 patients in total to identify a 11% reduction in felt pain. Given the negligible difference of 3.3 mm in VAS that this translates to, the pain's short duration, and the available safe methods to alleviate pain during fixed appliance treatment it would seem neither prudent nor easily ethically justifiable to perform such a trial. Such small trials might be meaningful in settings were the long-term survival of critically-ill patients can be increased by simple measures but would probably be regarded as research waste in the current setting. Assumptions about sample size calculation should ideally meet a fine balance among clinically relevant gains, careful management of research resources, and avoidance of unnecessarily putting patients in a clinical trial environment. Statement (C), which implies that when it comes to sample size more is always better is therefore not true.

Therefore, the mere fact that a sample size calculation has been performed does not necessarily mean that the trial is adequately powered. Critical appraisal of the assumptions made based on realistic knowledge of the field is needed in each case – provided of course that the sample size calculation is fully and transparently reported in the paper.

References

Cohen J. 1977. Statistical power analysis for the behavioral sciences (2nd ed.) New York: Academic Press.

Macfarlane TV. 2003. Sample size determination for research projects. J Orthod. 30(2):99—100.

Pandis N, Cobourne MT. 2013. Clinical trial design for orthodontists. J Orthod. 40(2):93—103.

Sandhu SS, Sandhu J. 2013. A randomized clinical trial investigating pain associated with superelastic nickel-titanium and multistranded stainless steel archwires during the initial leveling and aligning phase of orthodontic treatment. J Orthod. 40(4):276—285.

Figure Legend

Figure 1. Figure showing the relationship between statistical power (1-β; y axis) and total number of patients included in a two-group parallel trial for five different minimal clinically important differences with Cohen's d: 1.3, 1.0, 0.8, 0.5, and 0.3. All calculations are done assuming a pain response of 28.8 mm in VAS scale 24 hours after insertion for the control group, common standard deviations for the two groups, and α set at 5% for a t-test for independent samples. The pink dotted horizontal line denotes power of 80%.